Diagnosis of Bacterial Leaf Blight, Brown Spots, and Leaf Smut Rice Plant Diseases using Light GBM

G.R.I.L. Jayasooriya
University of Colombo School of Computing,
35, Reid Avenue, Colombo 7,
Sri Lanka

Samantha Mathara Arachchi
University of Colombo School of Computing,
35, Reid Avenue, Colombo 7,
Sri Lanka

ABSTRACT
Considering the human population, food is one of the major problems Sri Lanka might face in the near future. Rice is the most widely consumed food product and one of the extensively cultivated crops in Sri Lanka. Therefore, increasing the crop yield is one of the primary needs of the country. When rice crops are infected with diseases, it results in a loss of crops. Therefore, it is essential to identify the disease in the early stage of infection to prevent the damage that can be done. Disease identification could be challenging without a clear understanding. With the advancement of new technologies, researchers are interested in identifying paddy diseases through machine learning and image processing techniques to help farmers identify infectious diseases accurately.

It is difficult to observe the paddy leaf with the naked eye to diagnose the infected disease. In this research, an algorithm was developed to check whether the image contains different changes to the paddy leaf by considering the green colour pixels and their variance. OpenCV libraries have been used to develop the algorithm for feature extraction. Those features were used as attributes to the LightGBM algorithm to classify the disease images with over 80% accuracy.

General Terms
Algorithm, Machine Learning, Pattern recognition

Keywords
Decision Tree, Diagnosis, Diseases, Leaves, Light GBM, Open CV, Rice

1. INTRODUCTION
Paddy cultivation enjoys a prominent place in the agriculture industry in Sri Lanka because Rice is the principal food in the country[1]. Traditional varieties were cultivated throughout the country until the early 1960s [2]. But these varieties usually yield a low harvest, and to increase the yield, farmers use artificial fertilizers regardless of the possibility of the plants being attacked by pathogens[3]. Every year, farmers face a loss of yield and financial losses due to pests and diseases attacking the rice plants[3]. Many diseases affect the paddy yield[2][3]. The quality and quantity of rice are adversely affected due to infection caused by viruses, fungus, and bacteria[4][5]. It is therefore essential to identify these diseases to avoid their spreading and incurring huge losses. Traditional farmers can identify these diseases by observing the plants as they are well experienced. But novice farmers are unable to determine these diseases efficiently due to lack of experience.
Traditional farming methods are becoming obsolete with the advancement of new technologies. So the younger generation of farmers is more accustomed to using newly developed technologies to increase the yield. This incorporation of technology in farming can be termed smart farming.
Under smart farming, it is essential to diagnose plant diseases without human intervention to apply treatments to control the infection. The diagnosis requires a methodology rather than observation with the naked eye. Therefore, machine learning and image processing methods were suggested for field image diagnosis. The human generation would be smart enough to control their farms without physical attendance shortly, remotely.
It would be a challenge to observe symptoms of paddy plant diseases through an image sans any physical part. Colour changes of the image will be used as the main feature that can be extracted from an image. Images contain not only the paddy plant but also other objects. So, the paddy plant should be uniquely identified when image processing is applied to analyze paddy images to diagnose the disease. Paddy plant boundaries should be distinguished andother objects removed from the image.

2. LITERATURE REVIEW
Rice plants can be affected by viruses during the two stages of seed germination and seedling establishment. Due to the damage caused by diseases, the plant's stability, growth, and normal functioningare weakened[3]. Some of the noticeable damage to the plants include, gaining spotting and discoloration, gaining sterility, premature senescence of rice crop, and yellowing[6].Image processing can be used to identify these changes caused by disease infection.

Orillo et al. [7] proposed a methodology for monitoring crop Brown Spot, Rice Blast, and Bacterial Leaf diseases using the neural network concept with MATLAB. The neural network is trained through a database of diseased images. Image Enhancement, Image pre-process, and Image Segmentation were identified as steps to diagnose the rice blast, Bacterial Leaf Blight, and Rice Tungro using image processing.Mukherjee et al.[8] converted the original resized image into a gray image such that the pixels corresponding to the leaf image are the same. Then histogram was used in calculating the change in the pick value.

Mangla et al.[4] proposed Otsus' method[9] and vegetation segmentation for identifying the threshold values in images and use textual analysis to feature extraction and also Support Vector Machine (SVM) algorithm [10] has been proposed to classify the images considering the extracted features. According to Chaudhari et al.[11], the threshold of RGB images cannot be used to accurately identify the spot of disease from brown infected rice leaves and the entry on the 'H' component of the HSI colour model and 'Cr' component of YCbCrcolour model[12] sometimes identified as disease spots but not in all. However, the threshold on the' component

of the CIELAB[13]colour model allowed accurate detection of disease spots and is independent of background effects.

Mean value of the H, S, and, V of the disease, fraction covered by the disease on the leaf, standard deviation of the R, G, and B colour component of the disease and the arithmetic mean values for the R, G, and B colour components of the disease were extracted in this research. The number of iterations of the back-propagation algorithm[14]is trained until satisfactory results are achieved.

2.1 Presentation of Scientific Material

2.1.1 LightGBM

Gradient Boosting Decision Tree (GBDT) is a machine learning algorithm that is commonly used because of the veracity, interpretability and efficiency. LightGBM is a GBDT algorithm that incorporate methods like Gradient-based One-side Sampling (GOSS) and Exclusive Feature Bundling (EFB) in order to handle anextensive number of data instances and features. LightGBM was developed in April 2017 as a means to decrease the implementation time by a team from Microsoft. LightGBM can usuallysurpassXGBoost and SGB in terms of computational speed and memory with the help of GOSS and EFB[15]. The chief difference is that decision trees in LightGBM are expanded leaf-wise (Figure 2) without checking all of the previous leaves (Figure 1) for each new leaf, and all the attributes are sorted and grouped as bins. This is called histogram implementation. There are several notable advantages in LightGBM likehigher training speed, improved accuracy, and large-scale data handling is enabled, and GPU learning supported.

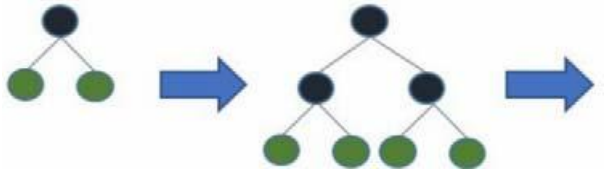

Figure 1 : Level wise tree growth XGBoost

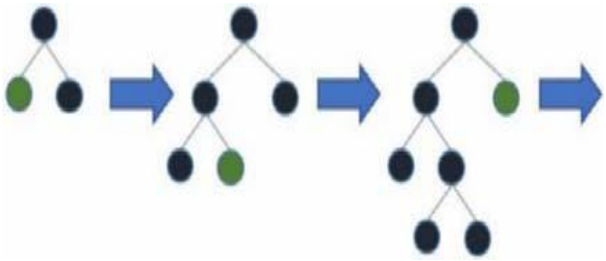

Figure 2 : Leaf wise tree growth LightGBM

2.1.2 K – mean

This is one of the widely used traditionalpartition-based clustering methods (Equation 1). K – mean algorithm is also one of the ten classical data mining algorithms. This algorithm basically cluster objects close to them by clustering K points in space. Until the best cluster results are obtained, thecentroid values in clusters are updated individually. The K-means algorithm portrays the clustering method based on the prototype function. The objective function of optimization is taken by considering the distance from the data point to the prototype. The adjustment rules of repititive operation are acquired by finding extreme values of functions. In order to keep the evaluation index to a minimum,this algorithm takes Euclidean distance as the similarity measure to find the optimal classification of a primary cluster center vector. As for a clustering criterion function, the error square sum criterion function is used. K-mean algorithm is considered efficient. However, the value of K should be given in advance. Further, the selection of K value is very strenuous to estimate. In most cases, the number of categories that the given data set should be divided into is undisclosed in advance[16].

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Equation 1 : K-mean Algorithm

2.1.3 Colour Spaces

This colour space describes colours (hue or tint) in terms of theirbrightness value and shade (saturation or amount of gray). The HSV colour wheel is illustrated as a cone or cylinder (Figure 3). HSV is abbreviated for Hue, Saturation and Value.Hue is given as a number from 0 to 360 degree. That is for red (which start at 0), yellow (beginning at 60), green (starting at 120), cyan (starting at 180), blue (starting at 240), and magenta (starting at 300). The amount of gray from zero percent to 100 percent in colour is depicted by the saturation. Value (or brightness) works combined with saturation. Itexpresses the brightness or intensity of the colourbetween zero percent to 100 percent (Figure 4). When selecting paint or ink people use the HSV colour space because itbest represents how people relate to colours than the RGB colour space. Another instance where the HSV colour wheel is used when generating high-quality graphics. Even though this is not widely known like RGB and CMYK, this is used in many high-end image editing software programs. When selecting an HSV colour, initially one of the available hues is picked, which most humans relate to colour. Then the shade and brightness value is adjusted. RGB and CMYK are defined proportionate to primary colours. But HSV is defined in correspondence with as to how humans perceive colour[17].

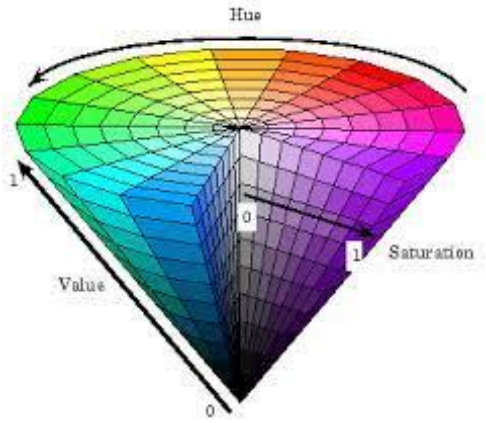

Figure 3 Illumination of HSV color space

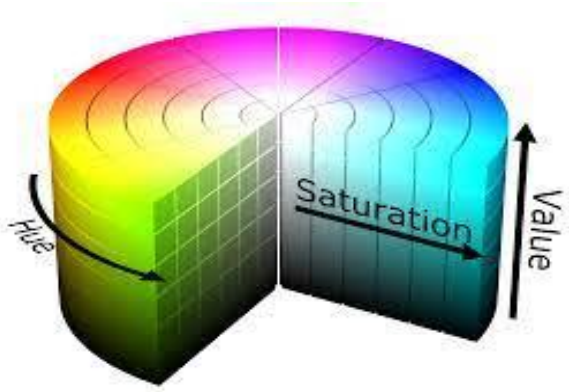

Figure 4: HSV color space

3. METHODOLOGY

Algorithm (Figure 5) was developed to isolate disease infected areas by accessing given image pixels' colours and the number of features have been identified referring the pixels which remained after removing background and leaf pixels. Those features were used as attributes to the LightGBM model. Inputs for the LightGBM algorithm were thirteen attributes extracted from discoloration pixels using python OpenCV libraries.

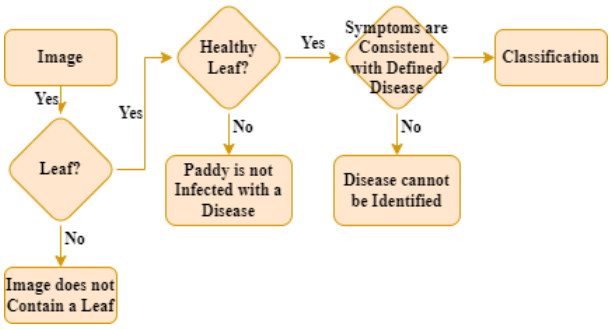

Figure 5: Flow of the model

3.1 Leaf Validation

Classification algorithms consider the similarities of the features of the given image with defined classes to check their features. In this case, when any image which does not contain a paddy leafis inserted it will also be classified as bacterial leaf blight, brown spot, or leaf smut after being compared with the most matching features of these classes. Therefore, it is essential to check whether paddy leaves are contained in the image uploaded to the website. It is hard to check whether there is a paddy leaf in the image. Therefore, a simple mathematical method was used to verify whether there are enough green colour pixels (Figure 6) in the image to have a paddy leaf and how close they are located in the image. Image background should include a non-green colour background to use this method.

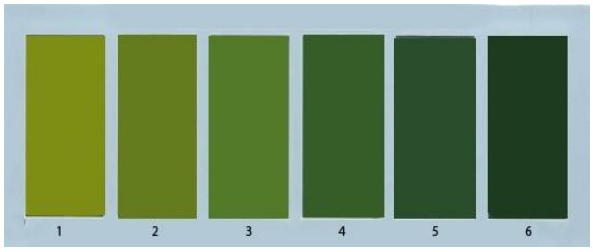

Figure 6: Leaf Color Chart

The Python OpenCV library has been used for the image validation part of this project. As a first step, all images should be the same size. Therefore, the uploaded image is resized to 150 width and 150 height to contain a fixed number of pixels. Moreover, the background of the image is eliminated to isolate leaf pixels. Green colour pixels have been used to achieve this task. RGB colour space wasn't supported for colour detection based on the RGB values. HSV (Hue, Value, and Saturate) colour space was identified as the most suitable colour space to convert numeric values to known colours. OpenCV reads the image as a BGR colour space, and hence after resizing the uploaded image, the image is converted into another three-colour space, respectively RGB, RGBA, and HSV. In RGBA, A for the alpha channel and this parameter can eliminate pixels from the image. HSV colour space image is used to determine whether pixel colour is in the green colour range. Lower HSV values for green colour were identified as (25, 52, 72) and higher values were identified as (126,255,255). That pixel in this range is identified from the HSV image and then applies bitwise operation with RGB image to remove green colour pixels. The highest and lowest green colour pixels are identified in each column, and those outside of lower and upper green colour pixels in each column are identified as background pixels. Then those pixels are eliminated from the image (Figure 7).

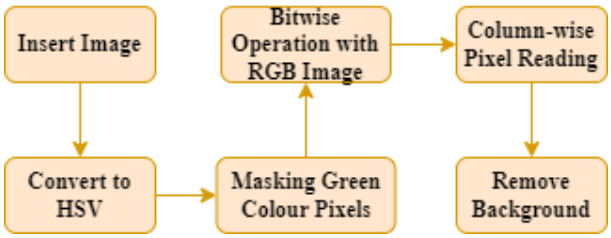

Figure 7: Steps of background pixels elimination

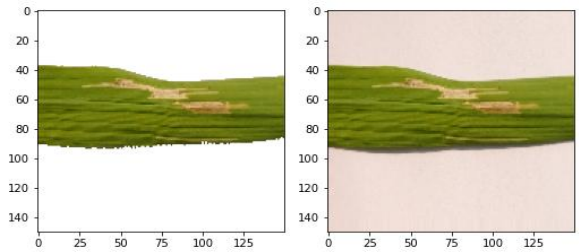

Figure 8: Paddy leaf image after and before removing the background

After eliminating background pixels from the image (Figure 8), the number of green colour pixels is counted and the standard divination of green colour pixels in the x and y-axis separately. Considering the number of green colour pixels in the image and their standard divination, the system defines whether the image consists of enough pixels using the number of pixels and how nearly all pixels are located using standard deviation (
Table ***1***).

Table 1: Leaf Validation Conditions

Number of pixels	Standard Deviation
less than 5,000	less than 20
less than 8,000 and greater than or equal to 5,000	less than 30
Greater than 8,000	less than 50

3.2 Disease Spots

Leaf border pixels and green colour pixels in the leaf have been identified in the leaf validation part. After validating an image containing a paddy leaf, discoloration areas should be determined to detect any disconsolation areas to have a disease in the paddy leaf. Green colour pixels are eliminated from the image the same as background pixels. After removing green colour pixels and background pixels, images only contain discoloration areas that can be used for paddy disease diagnosis (Figure 9). The remaining pixels with their colours can identify whether the leaf is healthy. According to the naked eye and open CV-based calculation observations, leaf discoloration areas should have at least one brown, yellow, black, and orange colour pixel (Figure 10). These characteristics check whether the paddy leaf is infected with one of the defined diseases. After removing background pixels from the image, leaf border pixels have been extracted before applying the K-mean algorithm (Figure 11). Python OpenCV library is used to predict the remaining pixel colour from the image. Image pixel colours can have various noises in the infected area's pixels and some unmatched colour pixels. Therefore, K-mean algorithms with k = 5 are applied in RGB colour space to segment the colour of the pixel with regard to the nearest pixel's colours. After applying the K-mean algorithm to image pixels (Figure 12), individual pixel colours are identified based on HSV colour space hue values. RGB colour space did not provide a good performance with pixel colour detection. Hence, images convert into HSV colour space and predict pixel colour using lower and higher value ranges of each colour. The number of brown, yellow, orange, red, and black pixels is calculated from the remaining pixels and checked whether there are enough pixels with these colours to have a disease that was described (Figure 12). Suppose the remaining pixels don't have enough pixels of the above colour, then it can identify whether that paddy leaf is infected with some other paddy diseases that this model cannot identify. Otherwise, images uploaded into a system classified as one of bacterial leaf blight, Leaf Smut, or Brown Spot infected leaf image using Constitutional Neural Network (CNN) are developed and tested using the dataset.

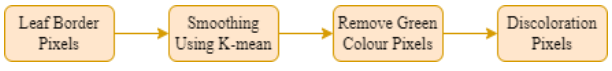

Figure 9: Discoloration area detection

3.3 Classification using LightGBM

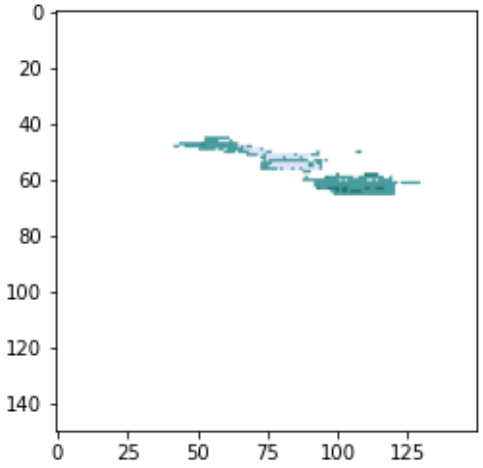

Figure 10: Infected areas pixels

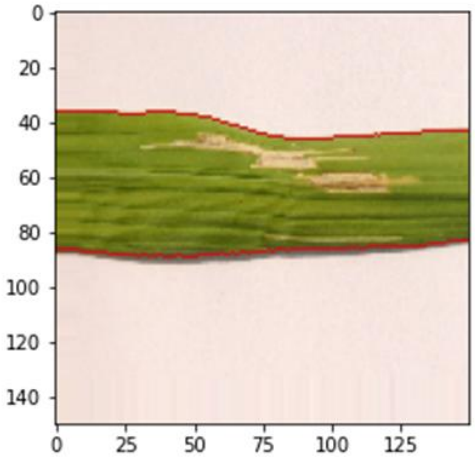

Figure 11: Paddy leaf border selection

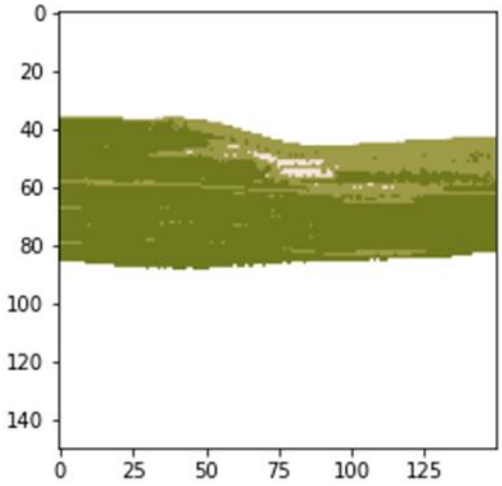

Figure 12: Paddy leaf after applying K-mean

A median filter was used to reduce noise in paddy leaf images. After applying the median filter to images, images were resized to 150,150 pixels using the python OpenCV algorithm. Considering the green colour pixels, background and paddy leaf boundaries define and remove background pixels from the image. After removing green colour pixels and background pixels from the image, only discoloration pixels can be identified as infected areas. Considering the row-wise pixel lengths of the remaining pixel located, thirteen features indicated below have been extracted from paddy leaf images (Table 2).

Table 2 : Attributes list for LightGBM

#	Attribute	Non – Null Count	Data Type
1	Shape	294 non-null	int64
2	Size	294 non-null	int64
3	Width	294 non-null	int64
4	Height	294 non-null	int64
5	No. of Orange Pixels	294 non-null	int64
6	No. of Brown Pixels	294 non-null	int64
7	No. of Black Pixels	294 non-null	int64
8	No. of Yellow Pixels	294 non-null	int64
9	No. of Gray Pixels	294 non-null	int64
10	No. of Red Pixels	294 non-null	int64
11	No. of Other Pixels	294 non-null	int64
12	Distance	294 non-null	int64
13	Line Count	294 non-null	int64

The colours of the remaining pixels were calculated using HSV colour space. The number of brown, yellow, black, red, orange, gray, and other colour pixels was calculated for each image and the number of pixels in the image (Figure 13). Image pixels have been accessed through X-axis, and the first and last pixels of each line were identified considering the neighbor pixel colours. Considering each continuous pixel in lines, the longest continuous pixel line was identified as the width of themost significant patch in the leaf. Moreover, a perpendicular line that goes through the centre of the longest width has been used to measure the height of the patch (Figure 14). The magnitude of width and height categorizes patch size into five categories.

Shapes of identifying the infected areas are hard to achieve because there were no exact shapes in images. Therefore, the width and height with their ratio have been used to define the shape of the most significant patch.

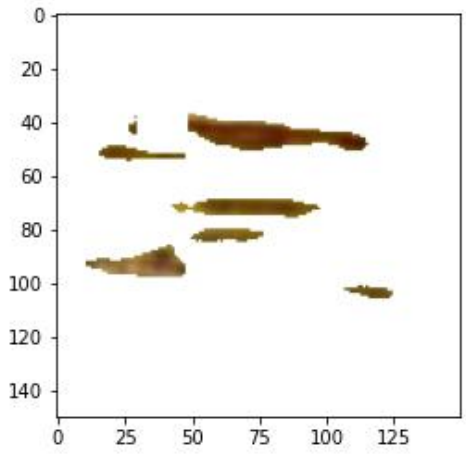

Figure 13: Discoloration pixel extraction

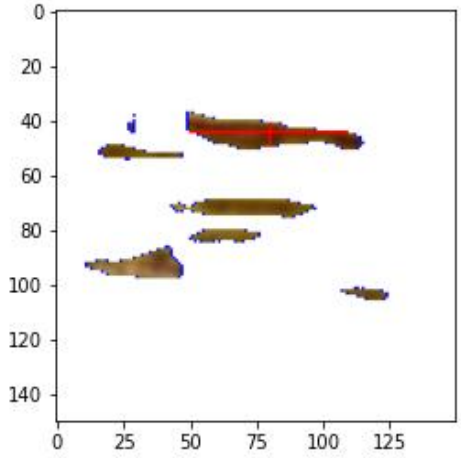

Figure 14: Feature extraction from the image

4. EVALUATION AND RESULTS

In this research, it is expected to find a way to diagnose rice leaf disease using naked eye observation attributes. Firstly it has to verify the leaf in the given image using several green colour pixels with their Standard Deviations. Considering the lack of capability to find pixel colours using RGB images, images were converted into HSV colour space for colour prediction. Discoloration areas of the image can be isolated after removing the background and green colour pixels from the image. Thirteen attributes have been calculated from the augmented paddy leaf image dataset and created aLightGBM algorithm with 81% accuracy (Table 3).

4.1 LightGBM matrix

Table 3: Confusion Matrix of LightGBM Model

#	Precision	Recall	F1-score	Support
0	0.85	0.92	0.88	25
1	0.77	0.71	0.74	28
2	0.81	0.81	0.81	36

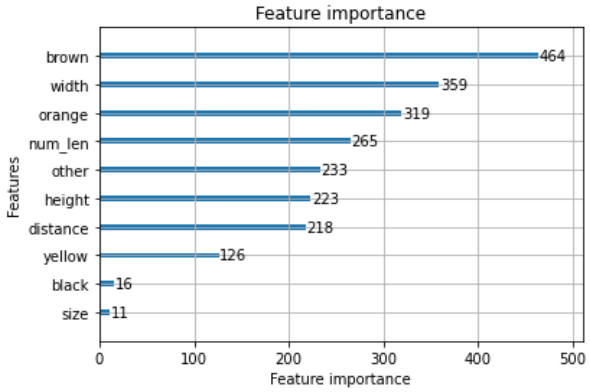

Figure 15: Feature importance

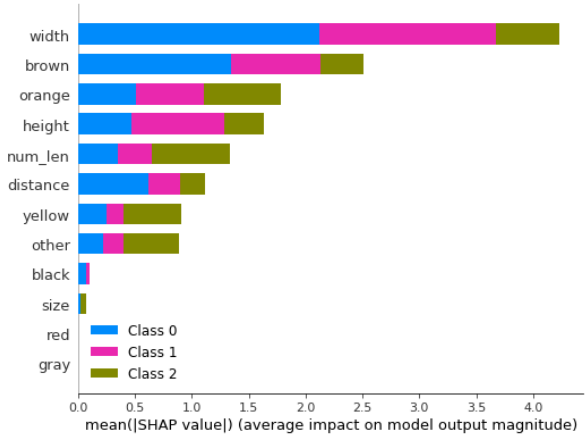

Figure 16: Average impact on the model output magnitude

The average impact of each class in the LightBGM model is represented in Figure 16, and the importance of individual features isillustrated in Figure 15. According to the diagram, a number of red and gray colour pixels do not provide enough support to the classification model. Attributes ‘width’ and ‘brown’ have a better-average impact on the model. Bacterial leaf blight, brown spot, and leaf smut are represented by class0, class1, and class2, respectively.

5. CONCLUSION AND FUTURE WORK

Algorithms to validate the paddy leaf from the image will be developed using leaf border pixels distribution with relevant positions. Here only the discoloration pixels’ colour from the picture as a feature is extracted. Colour feature alone could not be used to identify paddy disease. Therefore, discoloration areas’ shapes, size, and colour distribution over a single patch will be identified to provide a platform for diagnosing paddy leaf disease using their eye-catching features. The attributes mentioned have been extracted from Bacterial leaf blight, brown spot, and Leaf-smut images. The attributes with the LightGBM decision tree algorithm were provided over 80% accuracy for the paddy leaf diseases diagnosis model for the diseases mentioned above.

6. REFERENCES

[1] A. A. Gunawardana, "Agriculture sector performance in the Sri Lankan Economy: A systematic review and a Meta data analysis from 2012 to 2016.," 2018.

[2] R. Rambukwella and E. A. C. Priyankara, "Production and marketing of traditional rice varieties in selected districts in Sri Lanka: present status and future prospects," Hector Kobbekaduwa Agrarian Research and Training Institute, Colombo, Sri Lanka, 2016.

[3] S. N. Seneviratne, S. de and P. Jeyanandarajah, "RICE DISEASES -PROBLEM AND PROGRESS," Tropical Agricultural Research and Extension, 2004.

[4] D. N. Mangla, P. B. Raj and S. G. Hegde, "Paddy Leaf Disease Detection Using Image Processing and Machine Learning," 2019.

[5] P. K. Sethy, N. K. Barpanda, A. K. Rath and S. K. Behera, "Image Processing Techniques for Diagnosing Rice Plant Disease: A Survey," Procedia Computer Science , pp. 516-530, 2020.

[6] L. Nugaliyadde, N. Dissanayake and J. Mitrasena, "Advance in pest and disease management of rice in Sri Lanka : A review".

[7] J. W. Orillo, J. D. Cruz, L. Agapito, P. J. Satimbre and I. Valenzuela, "Identification of diseases in rice plants (oryza sativa) using back propagation Artificial Neural Network," in International Conference on Humanoid, Nanotechnology, Information Technology, Communication, and Control, Environment and Management (HNICEM), IEEE, Palawan, Philippines, 2014.

[8] M. Mukherjee, T. Pal and D. Samanta, "DAMAGED PADDY LEAF DETECTION USING IMAGE PROCESSING," 2010.

[9] C. Yu, C. Dian-ren, L. Yang and C. Lei, "Otsu’s thresholding method based on gray level-gradient two-dimensional histogram," in 2nd International Asia Conference on Informatics in Control, Automation and Robotics (CAR 2010), IEEE, Wuhan, China, 2010.

[10] W. Noble, "What is a support vector machine? Nat Biotechnol," p. 1565–1567, 2006.

[11] P. R. Chaudhari, N. Tamrakar, L. Singh, A. Tandon and D. Sharma, "Rice nutritional and medicinal properties: A review article 7".

[12] J. A. M. Basilio, G. A. Torres, S. Pérez, L. K. T. Medina and H. M. P. Meana, "Explicit Image Detection using YCbCr Space Color Model as Skin Detection," vol. 7, 2014.

[13] B. C. K. Ly, E. B. Dyer, J. L. Feig, A. L. Chien and S. D. Bino, "Research Techniques Made Simple: Cutaneous Colorimetry: A Reliable Technique for Objective Skin Color Measurement.," Journal of Investigative Dermatology, vol. 140, pp. 3-12, 2020.

[14] M. Buscema, "Back Propagation Neural Networks 38," 1998.

[15] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T. Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree".

[16] X. Zheng, Q. Lei, R. Yao, Y. Gong and Q. Yin, "Image segmentation based on adaptive K-means algorithm. J Image Video Proc," 2018.

[17] J. H. Bear, "Understanding HSV Color Model," ThoughtCo, 2017. [Online]. Available: www.thoughtco.com/what-is-hsv-in-design-1078068.